Rethinking Spatial Dimensions of Vision Transformers

Year: 2021 | Citations: 678 | Authors: Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe

Abstract

Vision Transformer (ViT) extends the application range of transformers from language processing to computer vision tasks as being an alternative architecture against the existing convolutional neural networks (CNN). Since the transformer-based architecture has been innovative for computer vision modeling, the design convention towards an effective architecture has been less studied yet. From the successful design principles of CNN, we investigate the role of spatial dimension conversion and its effectiveness on transformer-based architecture. We particularly attend to the dimension reduction principle of CNNs; as the depth increases, a conventional CNN increases channel dimension and decreases spatial dimensions. We empirically show that such a spatial dimension reduction is beneficial to a transformer architecture as well, and propose a novel Pooling-based Vision Transformer (PiT) upon the original ViT model. We show that PiT achieves the improved model capability and generalization performance against ViT. Throughout the extensive experiments, we further show PiT outperforms the baseline on several tasks such as image classification, object detection, and robustness evaluation. Source codes and ImageNet models are available at https://github.com/naver-ai/pit.